Working Paper Series, N. 13, September 2009

Structure Variability in Bayesian Networks

Marco Scutari
Department of Statistical Sciences
University of Padua
Italy

Abstract: The structure of a Bayesian network encodes most of the information about the probability distribution of the data, which is uniquely identified given some general distributional assumptions. Therefore it's important to study the variability of its network structure, which can be used to compare the performance of different learning algorithms and to measure the strength of any arbitrary subset of arcs.

In this paper we will introduce some descriptive statistics and the corresponding parametric and Monte Carlo tests on the undirected graph underlying the structure of a Bayesian network, modeled as a multivariate Bernoulli random variable.

Keywords: Bayesian network, bootstrap, multivariate Bernoulli distribution, structure learning algorithms.

Final version.

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Marco Scutari
tel: +39 049 827 4168
marco.scutari@stat.unipd.it

Structure Variability in Bayesian Networks

Marco Scutari
Department of Statistical Sciences
University of Padua
Italy

Abstract: The structure of a Bayesian network encodes most of the information about the probability distribution of the data, which is uniquely identified given some general distributional assumptions. Therefore it's important to study the variability of its network structure, which can be used to compare the performance of different learning algorithms and to measure the strength of any arbitrary subset of arcs.

In this paper we will introduce some descriptive statistics and the corresponding parametric and Monte Carlo tests on the undirected graph underlying the structure of a Bayesian network, modeled as a multivariate Bernoulli random variable.

Keywords: Bayesian network, bootstrap, multivariate Bernoulli distribution, structure learning algorithms.

1 Introduction

In recent years Bayesian networks have been successfully applied in several different disciplines, including medicine, biology and epidemiology (see for example Friedman *et al.* (2000) and Holmes and Jain (2008)). This has been made possible by a rapid evolution of structure learning algorithms, both constraint-based (from PC (Spirtes *et al.* 2001) to Grow-Shrink (Margaritis 2003) to IAMB (Tsamardinos *et al.* 2003) and its variants (Yaramakala and Margaritis 2005)) and score-based (from Greedy Equivalent Search (Chickering 2002) to genetic algorithms (Larrañaga *et al.* 1997)). The main goal in the development of these algorithms was the reduction of the number of either independence tests or score comparisons needed to learn the structure of the Bayesian network. Their correctness was proved assuming either very large sample sizes in relation to the number of variables (in the case of Greed Equivalent Search) or the absence of both false positives and false negatives (in the case of Grow-Shrink and IAMB). In most cases the characteristics of the learned networks were studied using a small number of reference data sets (Elidan 2001) as benchmarks, and differences from the true structure measured with descriptive measures such as Hamming distance (Jungnickel 2008).

This approach to model evaluation is not possible for real world data sets, as the true structure of their probability distribution is not known in advance. An alternative is provided by the use of either parametric or nonparametric bootstrap (Efron and

Tibshirani 1993). By applying the learning algorithm to a sufficiently large number of bootstrap samples it is possible to obtain confidence intervals and empirical probabilities for any feature of the network structure (Friedman *et al.* 1999), such as the presence or the composition of the Markov Blanket of a particular node. The fundamental limit in the interpretation of the results is that the "reasonable" level of confidence for thresholding depends on the data.
In this paper we propose a modified bootstrap-based approach for the inference on the structure of a Bayesian network. The undirected graph underlying the network structure is modeled as a multivariate Bernoulli random variable in which each component is associated with an arc. This assumption allows the derivation of both exact and asymptotic measures of the variability of the network structure or its parts.

2 Bayesian networks and bootstrap

Bayesian networks are graphical models where nodes represent random variables (the two terms are used interchangeably in this article) and arcs represent probabilistic dependencies between them (Korb and Nicholson 2004).
The graphical structure $\mathcal{G} = (\mathbf{V}, A)$ of a Bayesian network is a *directed acyclic graph* (DAG) which defines a factorization of the joint probability distribution of $\mathbf{V} = \{X_1, X_2, \ldots, X_v\}$, often called the *global probability distribution*, into a set of *local probability distributions*, one for each variable. The form of the factorization is given by the *Markov property*, which states that every random variable X_i directly depends only on its parents Π_{X_i}:

$$\mathsf{P}(X_1, \ldots, X_v) = \prod_{i=1}^{v} \mathsf{P}(X_i \,|\, \Pi_{X_i}) \qquad \text{(for discrete variables)} \tag{1}$$

$$f(X_1, \ldots, X_v) = \prod_{i=1}^{v} f(X_i \,|\, \Pi_{X_i}) \qquad \text{(for continuous variables).} \tag{2}$$

Therefore it's important to define confidence measures for specific features in the network structure, such as the presence of specific configurations of arcs. A related problem is the definition of a measure of variability for the network structure as a whole, both as an indicator of goodness of fit for a particular Bayesian network and as a criterion to evaluate the performance of a learning algorithm.
A possible solution for both these problems has been developed by Friedman *et al.* (1999) using bootstrap simulation, and modified by Imoto *et al.* (2002) to estimate the confidence in the presence of an arc (called *edge intensity*, and also known as *arc strength*) and its direction. This approach can be summarized as follows:

1. For $b = 1, 2, \ldots, m$
   (a) re-sample a new data set $\mathbf{D}^*_{\mathbf{b}}$ from the original data $\mathbf{D}$ using either parametric or nonparametric bootstrap.
   (b) learn a Bayesian network $\mathcal{G}_b$ from $\mathbf{D}^*_{\mathbf{b}}$.

2. Estimate the confidence in each feature f of interest as

$$\hat{\mathsf{P}}(f) = \frac{1}{m}\sum_{b=1}^{m} f(\mathcal{G}_b). \tag{3}$$

However, the empirical probabilities $\hat{\mathsf{P}}(f)$ are difficult to evaluate, because the distribution of $\mathcal{G}$ is unknown and the confidence threshold value depends on the data.

3 The multivariate Bernoulli distribution

Let $B_1, B_2, \ldots, B_k$, $k \in \mathbb{N}$ be Bernoulli random variables with marginal probability of success $p_1, p_2, \ldots, p_k$, that is $B_i \sim Ber(p_i)$, $i = 1, \ldots, k$. Then the distribution of the random vector $\mathbf{B} = [B_1, B_2, \ldots, B_k]^T$ over the joint probability space of $B_1, B_2, \ldots, B_k$ is a *multivariate Bernoulli random variable* (Krummenauer 1998b), denoted as $Ber_k(\mathbf{p})$. Its probability function is uniquely identified by the parameter collection

$$\mathbf{p} = \{p_I : I \subseteq \{1, \ldots, k\}, I \neq \varnothing\}, \tag{4}$$

which represents the *dependence structure* among the marginal distributions in terms of simultaneous successes for every non-empty subset I of elements of the random vector.
However, several useful results depend only on the first and second order moments of $\mathbf{B}$

$$\mathsf{E}(B_i) = p_i \tag{5}$$
$$\mathsf{VAR}(B_i) = \mathsf{E}(B_i^2) - \mathsf{E}(B_i)^2 = p_i - p_i^2 \tag{6}$$
$$\mathsf{COV}(B_i, B_j) = \mathsf{E}(B_i B_j) - \mathsf{E}(B_i)\mathsf{E}(B_j) = p_{ij} - p_i p_j \tag{7}$$

and the reduced parameter collection

$$\tilde{\mathbf{p}} = \{p_{ij} : i, j = 1, \ldots, k\}, \tag{8}$$

which is in fact used as an approximation of $\mathbf{p}$ in the generation random multivariate Bernoulli vectors in Krummenauer (1998a).

3.1 Uncorrelation and independence

Let's first consider a simple result that links covariance (and therefore correlation) and independence of two univariate Bernoulli variables.

Theorem 1. *Let B_i and B_j be two Bernoulli random variables. Then B_i and B_j are independent if and only if their covariance is zero:*

$$B_i \perp\!\!\!\perp B_j \iff \mathsf{COV}(B_i, B_j) = 0 \tag{9}$$

Proof. If B_i and B_j are independent then by definition

$$\mathsf{COV}(B_i, B_j) = p_{ij} - p_i p_j = \mathsf{P}(B_i = 1, B_j = 1) - \mathsf{P}(B_i = 1)\mathsf{P}(B_j = 1) = 0,$$

as $\mathsf{P}(B_i = 1, B_j = 1) = \mathsf{P}(B_i = 1)\mathsf{P}(B_j = 1)$.
If on the other hand we have that $\mathsf{COV}(B_i, B_j) = 0$, then

$$p_{ij} - p_i p_j = 0 \Rightarrow p_{ij} = p_i p_j \Rightarrow B_i \perp\!\!\!\perp B_j$$

which completes the proof. □

This theorem can be extended to multivariate Bernoulli random variables as follows.

Theorem 2. *Let* $\mathbf{B} = [B_1, B_2, \ldots, B_k]^T$ *and* $\mathbf{C} = [C_1, C_2, \ldots, C_l]^T$, $k, l \in \mathbb{N}$ *be two multivariate Bernoulli random variables. Then* $\mathbf{B}$ *and* $\mathbf{C}$ *are independent if and only if*

$$\mathbf{B} \perp\!\!\!\perp \mathbf{C} \iff \mathsf{COV}(\mathbf{B}, \mathbf{C}) = \mathbf{O} \tag{10}$$

where $\mathbf{O}$ *is the zero matrix.*

Proof. If $\mathbf{B}$ is independent from $\mathbf{C}$, then by definition every pair (B_i, C_j), $i = 1, \ldots, k$, $j = 1, \ldots, l$ is independent. Therefore the covariance matrix of $\mathbf{B}$ and $\mathbf{C}$ is

$$\mathsf{COV}(B_i, C_j) = c_{ij} = 0 \Longrightarrow \mathsf{COV}(\mathbf{B}, \mathbf{C}) = [c_{ij}] = \mathbf{O}$$

If conversely the covariance matrix $\mathsf{COV}(\mathbf{B}, \mathbf{C})$ is equal to the zero matrix, every pair (B_i, C_j) is independent as

$$c_{ij} = p_{ij} - p_i p_j = 0 \Longrightarrow p_{ij} = p_i p_j$$

This implies the independence of the random vectors $\mathbf{B}$ and $\mathbf{C}$, as their sigma-algebras

$$\sigma(\mathbf{B}) = \sigma(B_1) \times \ldots \times \sigma(B_k) \qquad \text{and} \qquad \sigma(\mathbf{C}) = \sigma(C_1) \times \ldots \times \sigma(C_l)$$

are functions of the sigma algebras induced by the two sets of independent random variables $B_1, B_2, \ldots, B_k$ and $C_1, C_2, \ldots, C_l$. □

The correspondence between uncorrelation and independence is identical to the analogous property of the multivariate Gaussian distribution (Ash 2000), and is closely related to the strong normality defined for orthogonal second order random variables in Loève (1977). It can also be applied to disjoint subsets of components of a single multivariate Bernoulli variable, as they are also distributed as multivariate Bernoulli random variables.

Theorem 3. *Let* $\mathbf{B} = [B_1, B_2, \ldots, B_k]^T$ *be a multivariate Bernoulli random variable; then every random vector* $\mathbf{B}^* = [B_{i_1}, B_{i_2}, \ldots, B_{i_l}]^T$, $\{i_1, i_2, \ldots, i_l\} \subseteq \{1, 2, \ldots, k\}$ *is a multivariate Bernoulli random variable.*

Proof. The marginal components of $\mathbf{B}^*$ are Bernoulli random variables, because $\mathbf{B}$ is multivariate Bernoulli. The new dependency structure is defined as

$$\mathbf{p}^* = \{p_{I^*} : I^* \subseteq \{i_1, \ldots, i_l\} \subseteq \{1, \ldots, k\}, I^* \neq \varnothing\},$$

and uniquely identifies the probability distribution of $\mathbf{B}^*$. □

Example 1. *Let's consider the trivariate Bernoulli random variable*

$$\mathbf{B} = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = \mathbf{B_1} + \mathbf{B_2} \qquad \textit{where} \qquad \mathbf{B_1} = \begin{bmatrix} 0 \\ B_2 \\ 0 \end{bmatrix} \qquad \textit{and} \qquad \mathbf{B_2} = \begin{bmatrix} B_1 \\ 0 \\ B_3 \end{bmatrix}.$$

Then the covariance matrix

$$\begin{aligned}
\mathsf{COV}(\mathbf{B_1}, \mathbf{B_2}) &= \mathsf{E}\left(\begin{bmatrix} 0 \\ B_2 \\ 0 \end{bmatrix} \begin{bmatrix} B_1 & 0 & B_3 \end{bmatrix}\right) - \mathsf{E}\left(\begin{bmatrix} 0 \\ B_2 \\ 0 \end{bmatrix}\right) \mathsf{E}\left(\begin{bmatrix} B_1 & 0 & B_3 \end{bmatrix}\right) \\
&= \mathsf{E}\left(\begin{bmatrix} 0 & 0 & 0 \\ B_1 B_2 & 0 & B_2 B_3 \\ 0 & 0 & 0 \end{bmatrix}\right) - \begin{bmatrix} 0 \\ p_2 \\ 0 \end{bmatrix} \begin{bmatrix} p_1 & 0 & p_3 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 & 0 \\ p_{12} & 0 & p_{23} \\ 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ p_1 p_2 & 0 & p_2 p_3 \\ 0 & 0 & 0 \end{bmatrix} = \\
&= \begin{bmatrix} 0 & 0 & 0 \\ p_{12} - p_1 p_2 & 0 & p_{23} - p_2 p_3 \\ 0 & 0 & 0 \end{bmatrix}
\end{aligned}$$

of the two components $\mathbf{B_1}$ *and* $\mathbf{B_2}$ *is equal to the zero matrix if and only if*

$$\begin{cases} p_{12} = p_1 p_2 \\ p_{23} = p_2 p_3 \end{cases} \Longrightarrow \{B_1 \perp\!\!\!\perp B_2, B_2 \perp\!\!\!\perp B_3\}$$

which in turn implies and is implied by $\mathbf{B_1} \perp\!\!\!\perp \mathbf{B_2}$.

3.2 Properties of the covariance matrix

The covariance matrix $\Sigma = [\sigma_{ij}]$, $i, j = 1, \ldots, k$ associated with a multivariate Bernoulli random vector has several interesting numerical properties. Due to the form of the central second order moments defined in formulas 6 and 7, the diagonal elements are bound in the interval

$$\sigma_{ii} = p_i - p_i^2 \in \left[0, \frac{1}{4}\right]. \tag{11}$$

The maximum is attained for $p_i = \frac{1}{2}$, and the minimum for both $p_i = 0$ and $p_i = 1$. For the Cauchy-Schwartz theorem (Ash 2000) then

$$0 \leqslant \sigma_{ij}^2 \leqslant \sigma_{ii}\sigma_{jj} \leqslant \frac{1}{16} \Longrightarrow |\sigma_{ij}| \in \left[0, \frac{1}{4}\right]. \tag{12}$$

The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ of Σ are similarly bounded, as shown in the following theorem.

Theorem 4. *Let* $\mathbf{B} = [B_1, B_2, \ldots, B_k]^T$ *be a multivariate Bernoulli random variable, and let* $\Sigma = [\sigma_{ij}]$, $i, j = 1, \ldots, k$ *be its covariance matrix. Let* λ_i, $i = 1, \ldots, k$ *be the eigenvalues of* Σ. *Then*

$$0 \leqslant \sum_{i=1}^{k} \lambda_i \leqslant \frac{k}{4} \tag{13}$$

and

$$0 \leqslant \lambda_i \leqslant \frac{k}{4}. \tag{14}$$

Proof. Since Σ is a real, symmetric, non-negative definite matrix, the eigenvalues λ_i are non-negative real numbers (Salce 1993); this proves the lower bound in both inequalities.
The upper bound in the first inequality holds because

$$\sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} \sigma_{ii} \leqslant \max_{\{\sigma_{ii}\}} \sum_{i=1}^{k} \sigma_{ii} = \sum_{i=1}^{k} \max \sigma_{ii} = \frac{k}{4},$$

as the sum of the eigenvalues is equal to the trace of Σ (Seber 2008). This in turn implies

$$\lambda_i \leqslant \sum_{i=1}^{k} \lambda_i \leqslant \frac{k}{4},$$

which completes the proof. □

These bounds define a convex set in $\mathbb{R}^k$, defined by the family

$$\mathcal{D} = \left\{ \Delta^{k-1}(c) : c \in \left[0, \frac{k}{4}\right] \right\} \tag{15}$$

where $\Delta^{k-1}(c)$ is the non-standard $k-1$ simplex

$$\Delta^{k-1}(c) = \left\{ (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k : \sum_{i=1}^{k} \lambda_i = c, \lambda_i \geqslant 0 \right\}. \tag{16}$$

3.3 Sequences of multivariate Bernoulli variables

Let's now consider a sequence of independent and identically distributed multivariate Bernoulli variables $\mathbf{B_1}, \mathbf{B_2}, \ldots, \mathbf{B_m} \sim Ber_k(\mathbf{p})$. The sum

$$\mathbf{S}_m = \sum_{i=1}^{m} \mathbf{B}_i \sim Bi_k(m, \mathbf{p}) \tag{17}$$

is distributed as a *multivariate Binomial random variable* (Krummenauer 1998b), thus preserving one of the fundamental properties of the univariate Bernoulli distribution. A similar result holds for the *law of small numbers*, whose multivariate

version states that a k-variate Binomial distribution $Bi_k(m, \mathbf{p})$ converges to a *multivariate Poisson distribution* $P_k(\mathbf{\Lambda})$:

$$\mathbf{S}_m \xrightarrow{d} P_k(\mathbf{\Lambda}) \qquad \text{as} \qquad m\mathbf{p} \to \mathbf{\Lambda}. \tag{18}$$

Both these distributions' probability functions, while tractable, are not very useful as a basis for explicit inference procedures. An alternative is given by the asymptotic *multivariate Gaussian distribution* defined by the *multivariate central limit theorem* (Ash 2000):

$$\frac{\mathbf{S}_m - m\mathsf{E}(\mathbf{B}_1)}{\sqrt{m}} \xrightarrow{d} N_k(\mathbf{0}, \Sigma). \tag{19}$$

The limiting distribution is guaranteed to exist for all possible values of $\mathbf{p}$, as the first two moments are bounded and therefore are always finite.

4 Inference on the network structure

Let $\mathcal{U} = (\mathbf{V}, E)$ be the undirected graph underlying the DAG $\mathcal{G} = (\mathbf{V}, A)$, defined as its unique biorientation (Bang-Jensen and Gutin 2009). Each edge $e \in E$ of $\mathcal{U}$ corresponds to the directed arcs in A with the same incident nodes, and has only two possible states (it's either present in or absent from the graph).
Then e_i, $i = 1, \ldots, |\mathbf{V} \times \mathbf{V}|$ is naturally distributed as a Bernoulli random variable

$$E_i = \begin{cases} e_i \in E & \text{with probability } p_i \\ e_i \notin E & \text{with probability } 1 - p_i \end{cases} \tag{20}$$

and every set $W \subseteq \mathbf{V} \times \mathbf{V}$ (including E) is distributed as a multivariate Bernoulli random variable $\mathbf{W}$ and identified by the parameter collection

$$\mathbf{p}_W = \{p_w : w \subseteq W, w \neq \varnothing\}. \tag{21}$$

The elements of $\mathbf{p}_W$ can be estimated via parametric or nonparametric bootstrap as in Friedman *et al.* (1999), because they are functions of the DAGs $\mathcal{G}_b$, $b = 1, \ldots, m$ through the underlying undirected graphs $\mathcal{U}_b = (V, E_b)$. The resulting empirical probabilities

$$\hat{p}_w = \frac{1}{m} \sum_{b=1}^{m} \mathbb{I}_{\{w \subseteq E_b\}}(\mathcal{U}_b), \tag{22}$$

in particular

$$\hat{p}_i = \frac{1}{m} \sum_{b=1}^{m} \mathbb{I}_{\{e_i \in E_b\}}(\mathcal{U}_b) \qquad \text{and} \qquad \hat{p}_{ij} = \frac{1}{m} \sum_{b=1}^{m} \mathbb{I}_{\{e_i \in E_b, e_j \in E_b\}}(\mathcal{U}_b), \tag{23}$$

can be used to obtain several descriptive measures and test statistics for the variability of the network's structure.

4.1 Interpretation of bootstrapped networks

Considering the undirected graphs $\mathcal{U}_1, \ldots, \mathcal{U}_m$ instead of the corresponding directed graphs $\mathcal{G}_1, \ldots, \mathcal{G}_m$ greatly simplifies the interpretation of bootstrap's results. In particular the variability of the graphical structure can be summarized in three cases according to the entropy (Cover and Thomas 2006) of the set of the bootstrapped networks:

- *minimum entropy*: all the networks learned from the bootstrap samples have the same structure, that is

$$E_1 = E_2 = \ldots = E_m = E. \tag{24}$$

  This is the best possible outcome of the simulation, because there is no variability in the estimated network. In this case the first two moments of the multivariate Bernoulli distribution are equal to

$$p_i = \begin{cases} 1 & \text{if } e_i \in E \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Sigma = \mathbf{O}. \tag{25}$$

- *intermediate entropy*: several network structures are observed with different frequencies m_b, $\sum m_b = m$. The first two sample moments of the multivariate Bernoulli distribution are equal to

$$\hat{p}_i = \frac{1}{m} \sum_{b\,:\,e_i \in E_b} m_b \qquad \text{and} \qquad \hat{p}_{ij} = \frac{1}{m} \sum_{b\,:\,e_i \in E_b, e_j \in E_b} m_b. \tag{26}$$

- *maximum entropy*: all $2^{|\mathbf{V}|}$ possible network structures appear with the same frequency, that is

$$\hat{\mathsf{P}}(\mathcal{U}_i) = \frac{1}{2^{|\mathbf{V}|}} \qquad i = 1, \ldots, 2^{|\mathbf{V}|}. \tag{27}$$

  This is the worst possible outcome because edges vary independently of each other and each one is present in only half of the networks (proof provided in appendix B):

$$p_i = \frac{1}{2} \qquad \text{and} \qquad \Sigma = \frac{1}{4} I_k. \tag{28}$$

4.2 Descriptive statistics of network's variability

Several functions have been proposed in literature as univariate measures of spread of a multivariate distribution, usually under the assumption of multivariate normality (see for example Muirhead (1982) and Bilodeau and Brenner (1999)). Three of them in particular can be used as descriptive statistics for the multivariate Bernoulli distribution:

- the *generalized variance*

$$\mathsf{VAR}_G(\Sigma) = \det(\Sigma). \tag{29}$$

- the *total variance* (called *total variation* in Mardia *et al.* (1979))

$$\mathsf{VAR}_T(\Sigma) = \mathrm{tr}(\Sigma). \tag{30}$$

- the squared *Frobenius matrix norm*

$$\mathsf{VAR}_N(\Sigma) = |||\Sigma - \frac{k}{4} I_k|||_F^2. \tag{31}$$

Both the *generalized variance* and the *total variance* associate high values of the statistic to unstable network structures, and are bounded due to the properties of the covariance matrix Σ. For the total variance it's easy to show that

$$0 \leqslant \mathsf{VAR}_T(\Sigma) = \sum_{i=1}^{k} \sigma_{ii} \leqslant \frac{1}{4} k \tag{32}$$

due to the bounds on the variances σ_{ii} in equation 11. The generalized variance is similarly bounded due to Hadamard's theorem on the determinant of a non-negative definite matrix (Seber 2008):

$$0 \leqslant \mathsf{VAR}_G(\Sigma) \leqslant \prod_{i=1}^{k} \sigma_{ii} \leqslant \left(\frac{1}{4}\right)^k. \tag{33}$$

They reach the respective maxima in the *maximum entropy* case and are equal to zero only in the *minimum entropy* case. The generalized variance is also strictly convex (the maximum is reached only for $\Sigma = \frac{1}{4} I_k$), but it is equal to zero if Σ is rank deficient. For this reason it's convenient to reduce Σ to a smaller, full rank matrix (let's say Σ^*) and compute $\mathsf{VAR}_G(\Sigma^*)$ instead of $\mathsf{VAR}_G(\Sigma)$.
The squared Frobenius norm on the other hand associates high values of the statistic to stable network structures. It can be rewritten in terms of the eigenvalues $\lambda_1, \ldots, \lambda_k$ of Σ as

$$\mathsf{VAR}_N(\Sigma) = \sum_{i=1}^{k} \left(\lambda_i - \frac{k}{4}\right)^2. \tag{34}$$

It has a unique maximum (in the *minimum entropy* case), which can be computed as the solution of the constrained minimization problem in $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_k]^T$

$$\min_{\mathcal{D}} f(\boldsymbol{\lambda}) = -\sum_{i=1}^{k} \left(\lambda_i - \frac{k}{4}\right)^2 \qquad \text{subject to} \qquad \lambda_i \geqslant 0, \sum_{i=1}^{k} \lambda_i \leqslant \frac{k}{4} \tag{35}$$

using the extended Lagrange multipliers methods (Nocedal and Wright 1999). It also has a single minimum in $\boldsymbol{\lambda}^* = [\frac{1}{4}, \ldots, \frac{1}{4}]$, which is the projection of $[\frac{k}{4}, \ldots, \frac{k}{4}]$ onto the set $\mathcal{D}$ and coincides with the *maximum entropy* case. The proof for these boundaries and the rationale behind the use of $\frac{k}{4} I_k$ instead of $\frac{1}{4} I_k$ are reported in appendix A.

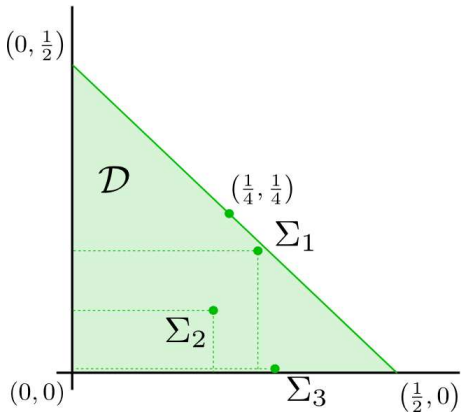

Figure 1: The covariance matrices Σ_1, Σ_2 and Σ_3 represented as functions of their eigenvalues in $\mathcal{D}$ (green). The points $(0, 0)$ and $(\frac{1}{4}, \frac{1}{4})$ correspond to the *minimum entropy* and *maximum entropy* cases.

The corresponding normalized statistics are:

$$\begin{aligned}
\overline{\mathsf{VAR}}_T(\Sigma) &= \frac{\mathsf{VAR}_T(\Sigma)}{\max_\Sigma \mathsf{VAR}_T(\Sigma)} = \frac{4\mathsf{VAR}_T(\Sigma)}{k} \\
\overline{\mathsf{VAR}}_G(\Sigma) &= \frac{\mathsf{VAR}_G(\Sigma)}{\max_\Sigma \mathsf{VAR}_G(\Sigma)} = \frac{\mathsf{VAR}_G(\Sigma)}{4^k} \\
\overline{\mathsf{VAR}}_N(\Sigma) &= \frac{\max_\Sigma \mathsf{VAR}_N(\Sigma) - \mathsf{VAR}_N(\Sigma)}{\max_\Sigma \mathsf{VAR}_N(\Sigma) - \min_\Sigma \mathsf{VAR}_N(\Sigma)} = \frac{k^3 - 16\mathsf{VAR}_N(\Sigma)}{k(2k-1)}.
\end{aligned}$$

All of them vary in the $[0, 1]$ interval and associate high values of the statistic to networks whose structure display a high variability across the bootstrap samples. Equivalently we can define

$$\begin{aligned}
\overline{\overline{\mathsf{VAR}}}_T(\Sigma) &= 1 - \overline{\mathsf{VAR}}_T(\Sigma) \\
\overline{\overline{\mathsf{VAR}}}_G(\Sigma) &= 1 - \overline{\mathsf{VAR}}_G(\Sigma) \\
\overline{\overline{\mathsf{VAR}}}_N(\Sigma) &= 1 - \overline{\mathsf{VAR}}_N(\Sigma)
\end{aligned}$$

which associate high values of the statistic to networks with little variability, and can be used as measures of distance from the *maximum entropy* case.

Example 2. *Let's consider three multivariate Bernoulli distributions* $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$ *with second order moments*

$$\Sigma_1 = \frac{1}{25}\begin{bmatrix} 6 & 1 \\ 1 & 6 \end{bmatrix}, \quad \Sigma_2 = \frac{1}{625}\begin{bmatrix} 66 & -21 \\ -21 & 126 \end{bmatrix}, \quad \textit{and} \quad \Sigma_3 = \frac{1}{625}\begin{bmatrix} 66 & 91 \\ 91 & 126 \end{bmatrix}.$$

The eigenvalues of Σ_1, Σ_2 *and* Σ_3 *are*

$$\boldsymbol{\lambda}_1 = \begin{bmatrix} 0.28 \\ 0.20 \end{bmatrix}, \qquad \boldsymbol{\lambda}_2 = \begin{bmatrix} 0.2121 \\ 0.095 \end{bmatrix}, \qquad \boldsymbol{\lambda}_3 = \begin{bmatrix} 0.3069 \\ 0.0003 \end{bmatrix}$$

and the values of the generalized variance, total variance and squared Frobenius matrix norm (both normalized and in the original scale) for the three covariance matrices are reported below.

	$\mathsf{VAR}_T(\Sigma)$	$\mathsf{VAR}_G(\Sigma)$	$\mathsf{VAR}_N(\Sigma)$	$\overline{\mathsf{VAR}}_T(\Sigma)$	$\overline{\mathsf{VAR}}_G(\Sigma)$	$\overline{\mathsf{VAR}}_N(\Sigma)$
Σ_1	0.48	0.056	0.1384	0.96	0.896	0.9642
Σ_2	0.3072	0.02016	0.2468	0.6144	0.32256	0.6752
Σ_3	0.3072	8.96×10^{-5}	0.2869	0.6144	0.00143	0.5682

4.3 Asymptotic inference

The limiting distribution of the descriptive statistics defined above can be derived by replacing the covariance matrix Σ with its unbiased estimator $\hat{\Sigma}$ and by considering the multivariate Gaussian distribution from equation 19. The hypothesis we are interested in is

$$H_0 : \Sigma = \frac{1}{4} I_k \qquad\qquad H_1 : \Sigma \neq \frac{1}{4} I_k, \tag{36}$$

which relates the sample covariance matrix with the one from the *maximum entropy* case.
For the total variance we have that (Muirhead 1982)

$$t_T = 4m \operatorname{tr}(\hat{\Sigma}) \dot{\sim} \chi^2_{mk}, \tag{37}$$

and since the maximum value of $\operatorname{tr}(\Sigma)$ is achieved in the *maximum entropy* case, the hypothesis in 36 assumes the form

$$H_0 : \operatorname{tr}(\Sigma) = \frac{k}{4} \qquad\qquad H_1 : \operatorname{tr}(\Sigma) < \frac{k}{4}. \tag{38}$$

Then the observed significance value is

$$\hat{\alpha}_T = \mathsf{P}(t_T \leqslant t_T^{oss}), \tag{39}$$

and can be improved with the finite sample correction

$$\tilde{\alpha}_T = \mathsf{P}\left(t_T \leqslant t_T^{oss} \,|\, t_T \in [0, mk]\right) = \frac{\mathsf{P}(t_T \leqslant t_T^{oss})}{\mathsf{P}(t_T \leqslant mk)} \tag{40}$$

which accounts for the bounds on $\mathsf{VAR}_T(\Sigma)$ from inequality 32.
For the generalized variance there are several possible asymptotic and approximate distributions:

- the Gaussian distribution defined in Anderson (2003)

$$t_{G_1} = \sqrt{m} \left(\frac{\det(\hat{\Sigma})}{\det(\frac{1}{4} I_k)} - 1 \right) \dot{\sim} N(0, 2k). \tag{41}$$

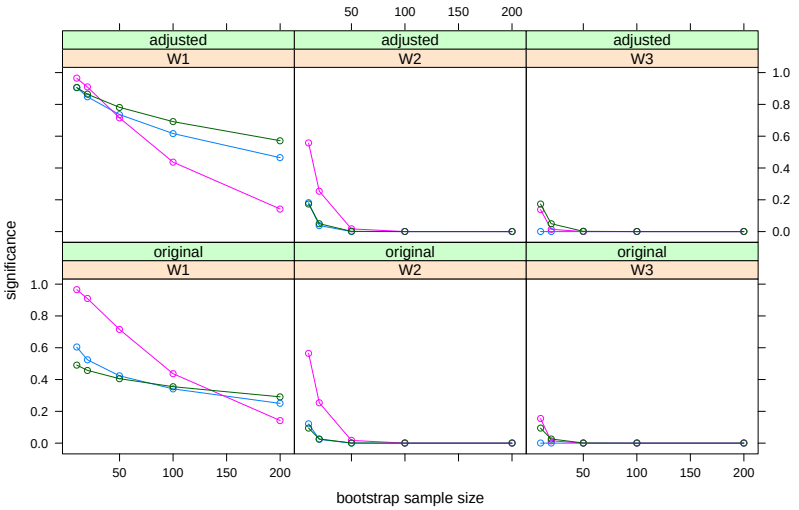

Figure 2: Asymptotic significance values of t_T (green), t_{G_2} (blue) and t_N (violet) for Σ_1, Σ_2 and Σ_3 from table 1.

- the Gamma distribution defined in Steyn (1978)

$$t_{G_2} = \frac{mk}{2}\sqrt[k]{\frac{\det(\hat{\Sigma})}{\det(\frac{1}{4}I_k)}} \dot{\sim} Ga\left(\frac{k(m+1-k)}{2}, 1\right). \tag{42}$$

- the saddlepoint approximation defined in Butler *et al.* (1992).

As before the hypothesis in 36 assumes the form

$$H_0 : \det(\Sigma) = \det\left(\frac{1}{4}I_k\right) \qquad H_1 : \det(\Sigma) < \det\left(\frac{1}{4}I_k\right). \tag{43}$$

The observed significance values for the Gaussian and Gamma distributions are

$$\hat{\alpha}_{G_1} = \mathsf{P}(t_{G_1} \leqslant t_{G_1}^{oss}) \qquad \hat{\alpha}_{G_2} = \mathsf{P}(t_{G_2} \leqslant t_{G_2}^{oss}) \tag{44}$$

and the respective finite sample corrections for the bounds on $\det(\Sigma)$ are

$$\tilde{\alpha}_{G_1} = \mathsf{P}\left(t_{G_1} \leqslant t_{G_1}^{oss} \,|\, t_{G_1} \in \left[-\sqrt{m}, 0\right]\right) = \frac{\mathsf{P}(t_{G_1} \leqslant t_{G_1}^{oss}) - \mathsf{P}(t_{G_1} \leqslant -\sqrt{m})}{\mathsf{P}(t_{G_1} \leqslant 0) - \mathsf{P}(t_{G_1} \leqslant -\sqrt{m})} \tag{45}$$

$$\tilde{\alpha}_{G_2} = \mathsf{P}\left(t_{G_2} \leqslant t_{G_2}^{oss} \,|\, t_{G_2} \in \left[0, \frac{mk}{2}\right]\right) = \frac{\mathsf{P}(t_{G_2} \leqslant t_{G_2}^{oss})}{\mathsf{P}(t_{G_2} \leqslant \frac{mk}{2})}. \tag{46}$$

The test statistic associated with the squared Frobenius norm is the test for the equality of two covariance matrices defined in Nagao (1973),

$$t_N = \frac{m}{2}\operatorname{tr}\left(\left[\hat{\Sigma}\left(\frac{1}{4}I_k\right)^{-1} - I_k\right]^2\right) = \frac{m}{2}\operatorname{tr}\left(\left[4\hat{\Sigma} - I_k\right]^2\right) \dot{\sim} \chi^2_{\frac{1}{2}k(k+1)}, \tag{47}$$

	$t_T(\Sigma)$				
	10	20	50	100	200
Σ_1	0.4911379	0.4576109	0.4054044	0.3549436	0.2912432
	0.906041	**0.863836**	**0.7814146**	**0.691495**	**0.571734**
Σ_2	0.0941934	0.0263308	0.0008529	0.0000038	1.09×10^{-10}
	0.1737661	**0.04970497**	**0.001644116**	**0.0000075**	$\mathbf{2.14 \times 10^{-10}}$
Σ_3	0.0941934	0.0263308	0.0008529	0.0000038	1.09×10^{-10}
	0.1737661	**0.04970497**	**0.001644116**	**0.0000075**	$\mathbf{2.14 \times 10^{-10}}$
	$t_{G_2}(\Sigma)$				
Σ_1	0.6039442	0.5242587	0.4231830	0.3411315	0.250054
	0.9052188	**0.8475223**	**0.7357998**	**0.6166961**	**0.4651292**
Σ_2	0.1214881	0.0235145	0.0002789	0.0000002	2.79×10^{-13}
	0.1820918	**0.03801388**	**0.000484961**	**0.00000045**	$\mathbf{5 \times 10^{-13}}$
Σ_3	3.13×10^{-10}	2.03×10^{-20}	9.82×10^{-51}	4.42×10^{-101}	1.26×10^{-201}
	$\mathbf{4.7 \times 10^{-10}}$	$\mathbf{3.28 \times 10^{-20}}$	$\mathbf{1.7 \times 10^{-50}}$	$\mathbf{7.99 \times 10^{-101}}$	$\mathbf{2.35 \times 10^{-201}}$
	$t_N(\Sigma)$				
Σ_1	0.9652055	0.9091238	0.7149371	0.4368392	0.1422717
	0.9645473	**0.9091083**	**0.7149371**	**0.4368392**	**0.1422717**
Σ_2	0.5649382	0.2537627	0.0170906	0.0001428	7.48×10^{-9}
	0.556708	**0.2536360**	**0.01709067**	**0.0001428399**	$\mathbf{7.48 \times 10^{-9}}$
Σ_3	0.1545514	0.0147960	0.0000085	2.37×10^{-11}	1.34×10^{-22}
	0.1385578	**0.01462880**	$\mathbf{8.5 \times 10^{-06}}$	$\mathbf{2.37 \times 10^{-11}}$	$\mathbf{1.34 \times 10^{-22}}$

Table 1: Asymptotic significance values of t_T, t_{G_2} and t_N for Σ_1, Σ_2 and Σ_3; the ones computed with the finite sample corrections are reported in bold.

because

$$\begin{aligned}\operatorname{tr}\left(\left[4\hat{\Sigma}-I_k\right]^2\right) &= \operatorname{tr}\left(\left[4\hat{\Sigma}-I_k\right]\left[4\hat{\Sigma}-I_k\right]\right) = 16\operatorname{tr}\left(\left[\hat{\Sigma}-\frac{1}{4}I_k\right]\left[\hat{\Sigma}-\frac{1}{4}I_k\right]\right) = \\ &= 16\operatorname{tr}\left(\left[U\Lambda U^H-\frac{1}{4}I_k\right]\left[U\Lambda U^H-\frac{1}{4}I_k\right]\right) = \\ &= 16\operatorname{tr}\left(U\left[\Lambda-\frac{1}{4}I_k\right]U^HU\left[\Lambda-\frac{1}{4}I_k\right]U^H\right) = 16\operatorname{tr}\left(\left[\Lambda-\frac{1}{4}I_k\right]^2\right) = \\ &= 16\sum_{i=1}^{k}\left(\lambda_i-\frac{1}{4}\right)^2 = 16|||\hat{\Sigma}-\frac{1}{4}I_k|||_F^2 \quad (48)\end{aligned}$$

where $U\Lambda U^H$ is the spectral decomposition of $\hat{\Sigma}$ (see appendix A for and explanation of the use of $\frac{1}{4}I_k$ instead of $\frac{k}{4}I_k$). The significance value for t_N is

$$\hat{\alpha}_N = \mathsf{P}(t_N \geqslant t_N^{oss}) \quad (49)$$

as the hypothesis in 36 becomes

$$H_0 : |||\Sigma-\frac{1}{4}I_k|||_F = 0 \qquad H_1 : |||\Sigma-\frac{1}{4}I_k|||_F > 0. \quad (50)$$

Unlike the previous statistics, Nagao's test displays a very good convergence speed, to the point that the finite sample correction for the bounds on the squared Frobenius matrix norm

$$\tilde{\alpha}_N = \mathsf{P}\left(t_N \geqslant t_N^{oss} \,|\, t_{G_1} \in [0, t_N^{max}]\right) = \frac{\mathsf{P}(t_N \geqslant t_N^{oss}) - \mathsf{P}(t_N > t_N^{max})}{\mathsf{P}(t_N \leqslant t_N^{max})} \tag{51}$$

is not appreciably better than the raw significance value.

Example 3. *Let's consider again the multivariate Bernoulli distributions* $\mathbf{W}_1$*,* $\mathbf{W}_2$*,* $\mathbf{W}_3$ *and their covariance matrices* Σ_1*,* Σ_2*,* Σ_3 *from example 2. The respective asymptotic significance values for the statistics* t_T*,* t_{G_1} *and* t_N *are reported in table 1.*

4.4 Monte Carlo inference and parametric bootstrap

Another approach to compute the significance values of the statistics $\mathsf{VAR}_T(\Sigma)$, $\mathsf{VAR}_G(\Sigma)$ and $\mathsf{VAR}_N(\Sigma)$ is again parametric bootstrap.
The multivariate Bernoulli distribution $\mathbf{W}_0$ specified by the hypothesis in 36 has a diagonal covariance matrix, so its components W_{0_i}, $i = 1, \ldots, k$ are uncorrelated. According to theorem 1 they are also independent, so the joint distribution of $\mathbf{W}_0$ is completely specified by the marginal distributions $W_{0_i} \sim Ber(\frac{1}{2})$. Therefore it's possible (and indeed quite easy) to generate observations from the null distribution and use them to estimate the significance value of the normalized statistics $\overline{\mathsf{VAR}}_T(\Sigma)$, $\overline{\mathsf{VAR}}_G(\Sigma)$ and $\overline{\mathsf{VAR}}_N(\Sigma)$ defined in section 4.2:

1. compute the value of test statistic T on the original covariance matrix Σ.
2. For $r = 1, 2, \ldots, R$.
   (a) generate m sets of k random samples from a $Ber(\frac{1}{2})$ distribution.
   (b) compute their covariance matrix Σ_r^*.
   (c) compute T_r^* from Σ_r^*
3. compute the Monte Carlo significance value as

$$\hat{\alpha}_R = \frac{1}{R} \sum_{r=1}^{R} \mathbb{I}_{\{x \geqslant T\}}(T_r^*). \tag{52}$$

This approach has two important advantages over the parametric tests defined in section 4.3:

- the test statistic is evaluated against the null distribution instead of its asymptotic approximation, thus removing any distortion caused by lack of convergence (which can be quite slow and problematic in high dimensions).
- each simulation r has a lower computational cost than the equivalent application of the structure learning algorithm to a bootstrap sample b. Therefore the Monte Carlo test can achieve a good precision with a smaller number of bootstrapped networks, allowing its application to larger problems.

	$\overline{\overline{\text{VAR}}}_T(\Sigma)$				
	10	20	50	100	200
Σ_1	0.569655	0.457109	0.129242	0.017416	0.000334
Σ_2	0.016834	0.000205	0	0	0
Σ_3	0.016834	0.000205	0	0	0
	$\overline{\overline{\text{VAR}}}_G(\Sigma)$				
Σ_1	0.784102	0.512839	0.14788	0.013678	0.000094
Σ_2	0.063548	0.000761	0	0	0
Σ_3	0.005909	0.000008	0	0	0
	$\overline{\overline{\text{VAR}}}_N(\Sigma)$				
Σ_1	0.743797	0.568819	0.239397	0.096544	0.019633
Σ_2	0.196996	0.037772	0.001018	0.000005	0
Σ_3	0.018292	0.000355	0	0	0

Table 2: Bootstrap significance values from parametric bootstrap for Σ_1, Σ_2 and Σ_3.

Example 4. *Let's consider the multivariate Bernoulli distributions* $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$ *from examples 2 and 3 one last time. The corresponding significance values for the three normalized statistics* $\overline{\overline{\text{VAR}}}_T(\Sigma)$, $\overline{\overline{\text{VAR}}}_G(\Sigma)$ *and* $\overline{\overline{\text{VAR}}}_N(\Sigma)$ *are reported in table 2 for various sizes of the bootstrap samples* ($m = 10, 20, 50, 100, 200$). *Each one have been computed from* $R = 10^6$ *covariance matrices generated from the null distribution. The code used for the simulation is reported in appendix C.*

5 Conclusions

In this paper we derived the properties of several measures of variability for the structure of a Bayesian network through its underlying undirected graph, which is assumed to have a multivariate Bernoulli distribution. Descriptive statistics, asymp-

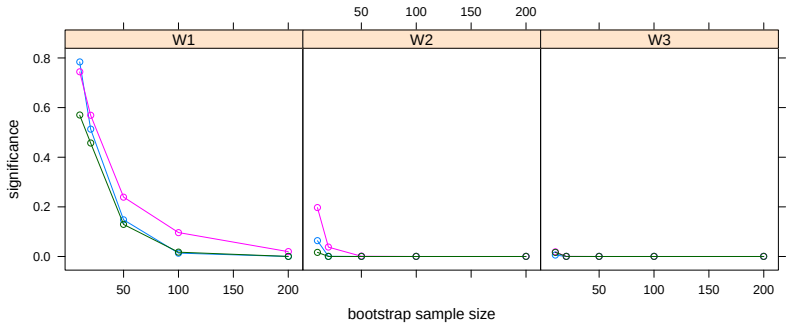

Figure 3: Monte Carlo significance values for the total variance (green), the generalized variance (blue) and the squared Frobenius matrix norm (violet) from table 2.

totic and Monte Carlo tests were developed along with their fundamental properties. They can be used to compare the performance of different learning algorithms and to measure the strength of any arbitrary subset of arcs.

Appendix

A Bounds on the squared Frobenius matrix norm

The squared Frobenius matrix norm of the difference between the covariance matrix Σ and the *maximum entropy* matrix $\frac{1}{4}I_k$ is

$$|||\Sigma - \frac{1}{4}I_k|||_F^2 = \sum_{i=1}^{k}\left(\lambda_i - \frac{1}{4}\right)^2. \tag{53}$$

Its unique global minimum is

$$|||\Sigma - \frac{1}{4}I_k|||_F^2 = 0 \tag{54}$$

for $\Sigma = \frac{1}{4}I_k$ due to the fundamental properties of the matrix norms (Salce 1993). However, it has a varying number of global maxima depending on the dimension k of Σ. They are the solutions of the constrained minimization problem

$$\min_{\mathcal{D}} f(\boldsymbol{\lambda}) = -\sum_{i=1}^{k}\left(\lambda_i - \frac{k}{4}\right)^2 \qquad \text{subject to} \qquad \lambda_i \geqslant 0, \sum_{i=1}^{k}\lambda_i \leqslant \frac{k}{4} \tag{55}$$

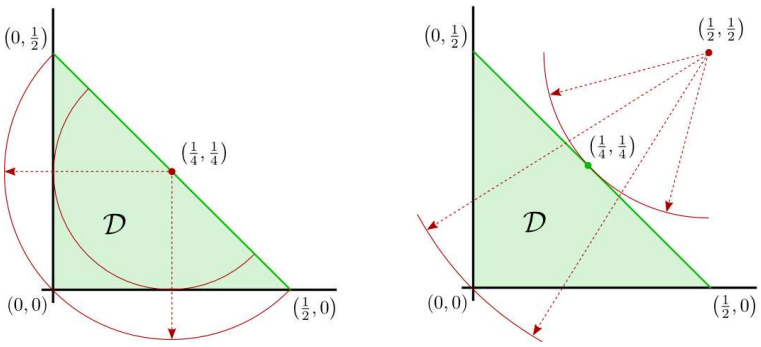

Figure 4: Squared Frobenius matrix norms from $\frac{1}{4}I_K$ (on the left) and $\frac{k}{4}I_k$ (on the right) in $\mathcal{D}$ for $k = 2$. The green area is the set $\mathcal{D}$ of the possible eigenvalues of Σ and the red lines are level curves.

and can be computed from the Lagrangian equation and its derivatives

$$\mathcal{L}(\boldsymbol{\lambda}, \mathbf{s}) = -\sum_{i=1}^{k}\left(\lambda_i - \frac{1}{4}\right)^2 - \sum_{i=1}^{k} s_i\lambda_i - s_{k+1}\left(\frac{k}{4} - \sum_{i=1}^{k}\lambda_i\right) \tag{56}$$

$$\frac{\delta}{\delta\lambda_i}\mathcal{L}(\boldsymbol{\lambda}, \mathbf{s}) = -2\lambda_i + \frac{1}{2} - s_i + s_{k+1} \tag{57}$$

$$\frac{\delta^2}{\delta^2\lambda_i}\mathcal{L}(\boldsymbol{\lambda}, \mathbf{s}) = -2, \qquad \frac{\delta^2}{\delta\lambda_i\delta\lambda_j}\mathcal{L}(\boldsymbol{\lambda}, \mathbf{s}) = 0 \tag{58}$$

where $\mathbf{s} = [s_1 \ldots, s_{k+1}]^T$ are the Lagrangian multipliers. This configuration of stationary points does not influence the results based on the asymptotic distribution of the multivariate Bernoulli distribution, but prevents any direct interpretation of quantities based on this difference in matrix norm as descriptive statistics.
On the other hand the difference in squared Frobenius norm

$$\mathsf{VAR}_N(\Sigma) = |||\Sigma - \frac{k}{4}I_k|||_F^2 = \sum_{i=1}^{k}\left(\lambda_i - \frac{k}{4}\right)^2 \tag{59}$$

has both a unique global minimum (because it's a convex function)

$$\min_{\mathcal{D}} \mathsf{VAR}_N(\Sigma) = \mathsf{VAR}_N\left(\frac{1}{4}I_k\right) = \sum_{i=1}^{k}\left(\frac{1}{4} - \frac{k}{4}\right)^2 = \frac{k(k-1)^2}{16} \tag{60}$$

and a unique global maximum

$$\max_{\mathcal{D}} \mathsf{VAR}_N(\Sigma) = \mathsf{VAR}_N(\mathbf{O}) = \sum_{i=1}^{k}\left(\frac{k}{4}\right)^2 = \frac{k^3}{16} \tag{61}$$

which correspond to the *minimum entropy* ($\boldsymbol{\lambda} = [0, \ldots, 0]$) and the *maximum entropy* ($\boldsymbol{\lambda} = [\frac{1}{4}, \ldots, \frac{1}{4}]$) covariance matrices respectively (see figure 4). However since $\frac{k}{4}I_k$ is not a valid covariance matrix for a multivariate Bernoulli distribution, $\mathsf{VAR}_N(\Sigma)$ cannot be used to derive any probabilistic result.

B Multivariate Bernoulli and the maximum entropy case

Let's first state a simple theorem on the probability of one and two edges in the *maximum entropy* case.

Theorem 5. *Let $\mathcal{U}_1, \ldots, \mathcal{U}_n$, $n = 2^{|\mathbf{V}|}$ be all possible undirected graphs with vertex set $\mathbf{V}$ and let*

$$\mathsf{P}(\mathcal{U}_k) = \frac{1}{n} \qquad\qquad k = 1, \ldots, n. \tag{62}$$

Let e_i and e_j, $i \neq j$ be two edges in $\mathbf{V} \times \mathbf{V}$. Then

$$\mathsf{P}(e_i) = \frac{1}{2} \qquad \textit{and} \qquad \mathsf{P}(e_i, e_j) = \frac{1}{4}. \tag{63}$$

Proof. The number of possible configurations of an undirected graph is given by the Cartesian product of the possible states of its edges, resulting in

$$|\{0,1\} \times \ldots \times \{0,1\}| = \left|\{0,1\}^{|\mathbf{V}|}\right| = 2^n \tag{64}$$

possible undirected graphs. Then edge e_i is present in

$$|\{0,1\} \times \ldots \times 1 \times \ldots \times \{0,1\}| = \left|1 \times \{0,1\}^{|\mathbf{V}|-1}\right| = 2^{n-1} \tag{65}$$

graphs and e_i and e_j are simultaneously present in

$$|\{0,1\} \times \ldots \times 1 \times 1 \times \ldots \times \{0,1\}| = \left|1^2 \times \{0,1\}^{|\mathbf{V}|-2}\right| = 2^{n-2} \tag{66}$$

graphs. Therefore

$$\mathsf{P}(e_i) = \frac{2^{n-1}\mathsf{P}(\mathcal{U}_k)}{2^n\mathsf{P}(\mathcal{U}_k)} = \frac{1}{2} \qquad \text{and} \qquad \mathsf{P}(e_i, e_j) = \frac{2^{n-2}\mathsf{P}(\mathcal{U}_k)}{2^n\mathsf{P}(\mathcal{U}_k)} = \frac{1}{4}. \tag{67}$$

□

Then the values of p_i and $\Sigma = [\sigma_{ij}]$ in equation 28 are indeed:

$$\mathsf{E}(e_i) = p_i = \frac{1}{2} \tag{68}$$

$$\mathsf{VAR}(e_i) = \sigma_{ii} = p_i - p_i^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \tag{69}$$

$$\mathsf{COV}(e_i, e_j) = \sigma_{ij} = p_{ij} - p_i p_j = \frac{1}{4} - \frac{1}{2} \cdot \frac{1}{2} = 0. \tag{70}$$

The fact that $\sigma_{ij} = 0$ for every $i \neq j$ also proves that the edges are independent according to theorem 1.

C R code for the parametric bootstrap simulation

The following R function has been used to compute the significance values in example 4.

```
biv.ber.sim = function(sigma, B, R, test) {

  if (test == "vart")
    FUN = function(lambda) 1/2 - sum(lambda)
  else if (test == "varg")
    FUN = function(lambda) 1/16 - prod(lambda)
  else if (test == "varn")
    FUN = function(lambda) sum((lambda - 1/4)^2)

  sim = matrix(0L, nrow = B, ncol = 2)
  tstar = numeric(R)

  s0 = eigen(sigma)$values
  t0 = FUN(s0)

  for (i in 1:R) {
```

```
    for (j in 1:B)
      sim[j, ] = rbinom(2, 1, 1/2)

    p = prop.table(table(factor(sim[, 1], levels = c(0,1)),
          factor(sim[, 2], levels = c(0,1))))

    sigmastar = matrix(
      c(sum(p[2,]) * (1 - sum(p[2,])),
        p[2,2] - sum(p[,2])*sum(p[2,]),
        p[2,2] - sum(p[,2])*sum(p[2,]),
        sum(p[,2]) * (1 - sum(p[,2]))),
      nrow = 2, ncol = 2, byrow = TRUE)

    sstar = eigen(sigmastar)$values
    tstar[i] = FUN(sstar)

  }

  return(length(tstar[tstar >= t0])/R)

}
```

The three covariance matrices Σ_1, Σ_2 and Σ_3 have been created in R with the following commands.

```
sigma1 = matrix(c(6, 1, 1, 6)/25, ncol = 2)
sigma2 = matrix(c(66, -21, -21, 126)/625, ncol = 2)
sigma3 = matrix(c(66, 91, 91, 126)/625, ncol = 2)
```

All the simulations have been performed on a Core Duo 2 machine with 1GB of RAM, with R 2.9.0 (R Development Core Team 2009) and an updated Debian GNU/Linux distribution.

D R code for the asymptotic inference

```
total.variance = function(sigma, b, adjusted = FALSE) {

  res = pchisq(4 * b * sum(diag(sigma)), 2 * b, lower.tail = TRUE)

  if (adjusted)
    res = res / pchisq(2 * b, 2 * b, lower.tail = TRUE)

  return(res)

}

generalized.variance = function(sigma, b, adjusted = FALSE) {

  res = pgamma(4 * b * sqrt(det(sigma)), b - 1, 1, lower.tail = TRUE)

  if (adjusted)
    res = res / pgamma(b, b - 1, 1, lower.tail = TRUE)

  return(res)

}

frobenius.norm = function(sigma, b, adjusted = FALSE) {

  res = pchisq(8 * b * sum((eigen(sigma)$values - 1/4 )^2), 3, lower.tail = FALSE)
```

```
  if (adjusted)
    res = (res - pchisq(b, 3, lower.tail = FALSE)) / pchisq(b, 3, lower.tail = TRUE)

  return(res)

}
```

References

Anderson TW (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 3rd edition.

Ash RB (2000). *Probability and Measure Theory*. Academic Press, 2nd edition.

Bang-Jensen J, Gutin G (2009). *Digraphs: Theory, Algorithms and Applications*. Springer, 2nd edition.

Bilodeau M, Brenner D (1999). *Theory of Multivariate Statistics*. Springer-Verlag.

Butler RW, Huzurbazar S, Booth JG (1992). "Saddlepoint Approximations for the Generalized Variance and Wilks' Statistic." *Biometrika*, **79**(1), 157 – 169.

Chickering DM (2002). "Optimal Structure Identification with Greedy Search." *Journal of Machine Learning Research*, **3**, 507–554.

Cover TA, Thomas JA (2006). *Elements of Information Theory*. Wiley.

Efron B, Tibshirani R (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Elidan G (2001). "Bayesian Network Repository." URL `http://www.cs.huji.ac.il/labs/compbio/Repository`.

Friedman N, Goldszmidt M, Wyner A (1999). "Data Analysis with Bayesian Networks: A Bootstrap Approach." In "Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)," pp. 206 – 215. Morgan Kaufmann.

Friedman N, Linial M, Nachman I (2000). "Using Bayesian Networks to Analyze Expression Data." *Journal of Computational Biology*, **7**, 601–620.

Holmes DE, Jain LC (eds.) (2008). *Innovations in Bayesian Networks: Theory and Applications*, volume 156 of *Studies in Computational Intelligence*. Springer.

Imoto S, Kim SY, Shimodaira H, Aburatani S, Tashiro K, Kuhara S, Miyano S (2002). "Bootstrap Analysis of Gene Networks Based on Bayesian Networks and Nonparametric Regression." *Genome Informatics*, **13**, 369–370.

Jungnickel D (2008). *Graphs, Networks and Algorithms*. Springer, 3rd edition.

Korb K, Nicholson A (2004). *Bayesian Artificial Intelligence*. Chapman and Hall.

Krummenauer F (1998a). "Efficient Simulation of Multivariate Binomial and Poisson Distributions." *Biometrical Journal*, **40**(7), 823 – 832.

Krummenauer F (1998b). "Limit theorems for multivariate discrete distributions." *Metrika*, **47**(1), 47 – 69.

Larrañaga P, Sierra B, Gallego MJ, Michelena MJ, Picaza JM (1997). "Learning Bayesian Networks by Genetic Algorithms: A Case Study in the Prediction of Survival in Malignant Skin Melanoma." In "Proceedings of the 6th Conference on Artificial Intelligence in Medicine in Europe (AIME'97)," pp. 261 – 272.

Loève M (1977). *Probability Theory*. Springer-Verlag, 4th edition.

Mardia KV, Kent JT, Bibby JM (1979). *Multivariate Analysis*. Academic Press.

Margaritis D (2003). *Learning Bayesian Network Model Structure from Data*. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. Available as Technical Report CMU-CS-03-153.

Muirhead RJ (1982). *Aspects of Multivariate Statistical Theory*. Wiley-Interscience.

Nagao H (1973). "On Some Test Criteria for Covariance Matrix." *The Annals of Statistics*, **1**(4), 700 – 709.

Nocedal J, Wright SJ (1999). *Numerical Optimization*. Springer-Verlag.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.r-project.org`.

Salce L (1993). *Lezioni sulle matrici*. Zanichelli.

Seber GAF (2008). *A Matrix Handbook for Stasticians*. Wiley-Interscience.

Spirtes P, Glymour C, Scheines R (2001). *Causation, Prediction and Search*. MIT Press.

Steyn HS (1978). "On Approximations for the Central and Noncentral Distribution of the Generalized Variance." *Journal of the American Statistical Association*, **73**(363), 670 – 675.

Tsamardinos I, Aliferis CF, Statnikov A (2003). "Algorithms for Large Scale Markov Blanket Discovery." In "Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference," pp. 376–381. AAAI Press.

Yaramakala S, Margaritis D (2005). "Speculative Markov Blanket Discovery for Optimal Feature Selection." In "ICDM '05: Proceedings of the Fifth IEEE International Conference on Data Mining," pp. 809–812. IEEE Computer Society, Washington, DC, USA.

Acknowledgements

Many thanks to Adriana Brogini, my Supervisor at the Ph.D. School in Statistical Sciences (University of Padova), for proofreading this article and giving many useful comments and suggestions. I would also like to thank Giovanni Andreatta and Luigi Salce (Full Professors at the Department of Pure and Applied Mathematics, University of Padova) for their help in the development of the constrained optimization and matrix norm applications respectively.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

Department of Statistical Sciences
University of Padua
Italy